Standing on the threshold of a virtual library

Ian H. Witten

Department of Computer Science
University of Waikato
Hamilton, New Zealand
ihw@cs.waikato.ac.nz

1 Introduction

Physical libraries have been around for more than twenty-five centuries. Long before even the great Alexandrian Library, the Assyrian king Assurbanipal (668-626 B.C.) established a comprehensive, well-organized collection of some tens of thousands of clay tablets. In an early statement of library policy, an Alexandrian librarian was reported as being "anxious to collect, if he could, all the books in the inhabited world, and, if he heard of, or saw, any book worthy of study, he would buy it"—and, two millenia later, Thompson (1977) formulated this as a self-evident principle of librarianship: *It is a librarian's duty to increase the stock of his library.* When asked how large a library should be, librarians answer "bigger. And with provision for further expansion."

Only recently has the Alexandrian principle begun to be questioned. In 1974, following a 10-year building boom unprecedented in library history, the *Encyclopaedia Britannica* noted that "even the largest national libraries are ... doubling in size every 16 to 20 years," and gently warned that "such an increase can hardly be supported indefinitely." A collection of essays published in 1976 entitled *Farewell to Alexandria: Solutions to space, growth, and performance problems of libraries* dwells on the problems that arise when growth must end (Gore, 1976). Sheer limitation of space have forced librarians to rethink their principles. Now they talk about "aggressive weeding" and "culling," "no-growth" libraries, the "optimum size for collections," and some even ask "could smaller be better?"[1] The notion of focused collections is replacing the Alexandrian model that the ideal library

[1] In a striking example of *very* agressive weeding, the library world was rocked in 1996 by allegations that the San Francisco Public Library had surreptitiously dumped 200,000

is vast and ever-growing. The notion of service to library users is replacing the idea of a library as a storehouse of the world's knowledge. Perhaps this movement is reinforced by the experience of the World Wide Web, which amply illustrates the anarchy and chaos that inevitably result from sustained exponential growth. The events of the last quarter century have even shaken librarians' confidence in the continued existence of the traditional library. Defensive tracts with titles like *Future libraries: dreams, madness and reality* deride "technolust" and the empty promises of the technophiles (Crawford and Gorman, 1995).

Let us, for a moment at least, give an ear to the technophiles. Over sixty years ago, science fiction writer H.G. Wells was promoting the concept of a "world brain" based on a permanent world encyclopedia which "would be the mental background of every intelligent [person] in the world. It would be alive and growing and changing continually under revision, extension and replacement from the original thinkers in the world everywhere," and he added sardonically that "even journalists would deign to use it" (Wells, 1938). Eight years later, Vannevar Bush, the highest-ranking scientific administrator in the U.S. war effort, invited us to "consider a future device for individual use, which is a sort of mechanized private file and library ... a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility" (Bush, 1947). Fifteen years later, J.C.R. Licklider, head of the U.S. Department of Defense's Information Processing Techniques Office, envisioned that human brains and computing machines would be coupled together very tightly, and imagined this to be supported by a "network of 'thinking centers' that will incorporate the functions of present-day libraries together with anticipated advances in information storage and retrieval" (Licklider, 1960). Thirty-five years later we became accustomed to hearing similar pronouncements from the U.S. Presidential office rising above the road noise of the information superhighway.

But where *is* the virtual library? To paraphrase the dictionary definition, something is "virtual" if it exists in essence or effect though not in actual fact, form, or name. A virtual library is a library for all practical purposes, but a library without walls—or books.

In truth, a "virtual" representation of books has been at the very center of libraries right from the very beginning: the catalog. Even before Alexandria, libraries were arranged by subject and had catalogs that gave the title

books, or 20% of its collections, into landfills, because its new building, though lavishly praised by architecture critics, was too small for all the books.

of each work, the number of lines, the contents, and the opening words. In 240 B.C. an index was produced to provide access to the books in the Alexandrian library that was a classified subject catalog, a bibliography and a biographical dictionary all in one (Thompson, 1977).

A library catalog is a complete model that represents, in a predictable manner, the universe of books in the library: it furnishes a clear view of the content of the collection. Catalogs provide a summary of, if not a surrogate for, library contents. Today we call this "metadata": data about data. As Bowker (1883), himself a late 19th Century librarian, rather smugly wrote, "librarians classify and catalog the records of ascertained knowledge, the literature of the whole past. ... In this busy generation ... the librarian makes time for his fellow mortals by saving it. ... And this function of organizing, of indexing, of time-saving and thought-saving, is associated peculiarly with the librarian of the nineteenth century."

Other essential aids to information-seeking in libraries are published bibliographies and indexes. Like catalogs, these are virtual representations—metadata—and they provide the traditional means of gaining access to journal articles, government documents, microfiche and microfilm and special collections (Mann, 1993).

The information in library catalogs and bibliographies can be divided into two kinds: the first having reference to the contents of books; the second treating their external character and the history of particular copies. Intellectually, only the content of a library—the information contained in it—seems important. But the strong visceral element of books cannot be neglected—and is often cited as a reason why book collections will never become "virtual." Bibliophiles love books as much for the statements they make as objects as for the statements they contain as text. Beautiful books are highly prized for their splendid illustrations, for colored impressions, for heavily-decorated illuminated letters, for being printed on uncommon paper, or uncommon materials, for their unusual bindings. In the library in the castle of Königsburg are twenty books bound in silver, richly adorned with large and beautifully-engraved gold plates. Whimsical bindings abound: a London bookseller had Fox's *History of King James II* bound in fox-skin, and history provides many examples of books bound in human skin.[2] (Those

[2]It is hard to resist just one macabre example: a book in the Boston Athenaeum's collection. "James Allen, alias George Walton, was a burglar, bank robber, horse thief and highwayman when, in 1833, he attacked John Fenno Jr. on the Salem, Massachusetts, Turnpike with intent to rob. Fenno resisted his attacker and was shot, but saved by a suspender buckle. Allen fled, was caught and sent to prison where he wrote a boastful autobiographical account of his life of crime called *The Highwayman*. Admiring Fenno's

who feel nauseous may find this the best argument of all for virtual libraries!)

Catalogs and bibliographies comprise metadata: virtual information about books. In the kind of virtual library conceived by the visionary technophiles quoted above, the very concept of the book as an individual physical entity is at risk. However, it need not be: surrogates can substitute for physical books. A picture of the cover may be displayed as a "tangible"—or at least memorable—emblem of the physical book itself. A user may even be able to direct an avatar that strolls through a library's bookstacks to browse the collection using graphical techniques of virtual reality. But it is unlikely, perhaps inappropriate, that readers will "love" simulated books the way that bibliophiles love real ones, and eventually surrogates may become anacronistic and fade away. For what really matters in libraries is knowledge.

The primary portal into the virtual collection is the same as the primary portal into any physical collection: the catalog. The difference is that it is not just the catalog that is virtual, it is the books themselves. It is possible to create a "virtual collection" of material simply by gathering together a single catalog, and using it to reference the physical location of the data anywhere in the world. Indeed, networks exist where it does not make sense to talk about the location of the data at all.[3] These offer unprecedented safeguards against censorship, but—by the same token—pose unprecedented problems for copyright and intellectual property protection.

Virtuality offers far more than merely facilitating the act of gathering together information collections. Libraries are as much about *access* as they are about collecting. Public access to libraries goes right back to the clay tablet libraries of Assyria—although ancient libraries were only useful to the small minority of people who could read, and moreover were accessible only within stringent limitations imposed by social conditions. Medieval monastic and university libraries held chained copies of books in public reading areas and other copies that were available for loan—although very substan-

---

bravery, he asked that Fenno be given a copy of his book bound in the author's skin.

On July 17, 1837 upon Allen's death, Massachusetts General Hospital 'accepted his body for anatomical and pathological studies' and removed enough skin to provide the covering of his book. Bookbinder Peter Low treated the skin to look like gray deerskin and edged it with gold tooling. It is embossed with the Latin inscription 'Hic Liber Waltonis Cute Compactus Est' (This book by Walton is bound in his own skin)."—Kruse (1994)

[3]For example, Freenet allows information to be distributed over the Internet in an efficient, completely decentralized, manner. There is no person, computer, or organisation that is essential to its operation. It "learns" to route requests more efficiently, automatically mirrors popular data, makes network flooding almost impossible, and moves data to where it is in greatest demand (Clarke, 1999).

tial security was often demanded for each volume borrowed. The public library movement took hold in the U.K. and the U.S. in the 19th Century. Still, the libraries of the day had bookstacks that were closed to the public: patrons perused the catalog, chose their books, and they were handed out over the counter. In continental Europe, most libraries still operate with closed stacks. However, progressive 20th Century librarians came to realize the advantage of allowing readers to browse among the shelves and make their own selection, and the idea of open-access libraries became widely adopted in the U.K. and U.S. Indeed, Thompson (1977) regards open access as a particular contribution of the American library movement, and adds: "more than anything else, it marks the fulfilment of the principle of free access to the contents of libraries by all: the symbolic snapping of the links of the chained book." He goes on to state this as another principle of librarianship: *Libraries are for all.*

Today, we stand on the threshold of the virtual library—often called the "digital" or "electronic" library. Libraries are society's repositories for knowledge. Physical libraries began in an era where agriculture was humankind's greatest preoccupation, experienced a resurgence with the invention of printing in the Renaissance, and really began to flourish when the industrial revolution prompted a series of inventions that mechanized the printing process—the steam press, for example. The information revolution not only supplies the underlying enabling technology for virtual libraries, but has provoked an unprecedented thirst for storing, organizing, and accessing information. If information is the currency of the knowledge economy, virtual libraries will be the banks where it is invested.

Virtual libraries are of the utmost strategic importance in any modern economy. They will likely figure amongst the most important and influential institutions of the 21st Century. Recognizing their importance, many countries have initiated large-scale digital library projects (Lesk, 1997). In the mid-1990s the Digital Library Initiative was established in the U.S., generously funded by NSF, NASA, and ARPA. In the U.K., a national review called for "a sea-change in the way institutions plan and provide for the information of those working within them" (Follett, 1993), prompting a national Electronic Libraries (E-Lib) program. Other countries in Europe and the Pacific Rim have followed suit.

What will it be like to work in a virtual library? Will it feel like a conventional library, but more computerized, more networked, more international, more all-encompassing, more convenient? I believe the answer is no: it will feel qualitatively different. Not only will the library be with you on your desktop (or at the beach, or in the plane), but information workers

will work "inside" the library in a way that is quite unlike how they operate at present. It's not just that knowledge and reference services will be fully portable, operating round the world, around the clock, throughout the year, freeing library patrons from geographic and temporal constraints—important and liberating as these are. It's that when you read, you will be constantly surrounded by the related literature: not only will it always be at your fingertips, but it will effortlessly reorganize itself to track what you are attending to, moment by moment. It's that when new knowledge is created it will be fully contextualized and both sited within and cited by existing literature right from its conception.

This paper explores the notion of a virtual library. We ground the discussion with a necessarily cursory review of the development of physical libraries, in the next section, and summarize the short history and current state of the Internet as a global information resource, in the following one. This establishes the background for virtual libraries from the library and technology viewpoints. Next we define what is meant by a "virtual library," culling definitions from the literature to provide a simple and succinct characterization of the term. The following sections give examples of prototype virtual library collections. We stress the international aspect of digital libraries and the enormous increase in potential user access that they promote. Both full-text searching and metadata-directed browsing are easy to implement and provide flexible ways of getting at the information you need. Then we take a look at a public-domain, open-source software system that facilitates the building and maintenance of such collections. Finally, we return to our central theme: that the promise of virtual libraries has arrived in the nick of time, when physical libraries are beginning to crack under the inescapable strains of sustained exponential growth.

2 The first two and a half millenia

We begin with the fabled library of Alexandria in Egypt—although, as we have seen, it was not the first. Created in 300 B.C., it grew at a phenomenal rate, and, according to legend, contained some 200,000 volumes within ten years. Working in the acquisitions department in those days was pretty exciting. During a famine, the king refused to sell corn to the Athenians unless he received in pledge the original manuscripts of some leading authors. The manuscripts were diligently copied and the copies returned to the owners, while the originals went into the library. By far the largest single acquisition occurred when Mark Antony stole the rival library of Pergamum and gave

it lock, stock, and barrel—200,000 volumes—to Cleopatra as a love token; she passed it over to Alexandria for safe keeping. By the time Julius Caesar fired the harbor of Alexandria in 47 B.C., the library had grown to 700,000 volumes. More than two thousand years would pass before any other library would attain this size, notwithstanding technological innovations such as the printing press.

Tragically, the Alexandrian library was destroyed. Much remained after Caesar's fire, but this was wilfully laid waste (according to the Moslems) by Christians in 391 A.D. or (according to the Christians) by Moslems in 641 A.D. In the Arab conquest, Amru, the captain of Caliph Omar's army, would apparently have been willing to spare the library, but the fanatical Omar is said to have disposed of the problem of information explosion with the immortal words, "If these writings of the Greeks agree with the Koran they are useless, and need not be preserved; if they disagree they are pernicious, and ought to be destroyed."

Moving ahead a thousand years, let us peek at what was happening in a major university library a century or two after Gutenberg's invention of the movable-type printing press around 1450. Trinity College, Dublin, one of the oldest universities in Western Europe, was founded in 1592 by Queen Elizabeth I. In 1600 the library contained a meager collection of thirty books and ten manuscripts. However, this grew rapidly, by several thousand, when two of the Fellows mounted a shopping expedition to England, and by a further ten thousand when the library received the personal collection of Archbishop Ussher, a renowned Irish man of letters, on his death in 1661. Another great event in the development of the library occurred in 1801, when an Act was passed by the British Parliament decreeing that a copy of every book printed in the British Isles was to be donated to the Trinity College Library.

There were no journals in Ussher's collection. The first scholarly journals appeared just after his death: the *Journal des Sçavans* began in January 1665 in France, and the *Philosophical Transactions of the Royal Society* of the Royal Society began in March 1665 in England. These two have grown, hydra-like, into some fifty thousand scientific journals today.

In the 18th Century, the technology of printing really took hold. For example, more than thirty thousand titles were published in France during a sixty-year period in the mid 1700s. The printing press that Gutenberg had developed in order to make the Bible more widely available became the vehicle for disseminating the European Enlightenment—an emancipation of human thinking from the weight of authority of the church—some three hundred years later.

In the United States, President John Adams created a reference library for Congress when the seat of government was moved to the new capital city of Washington in 1800. He began by providing $5,000 "for the purchase of such books as may be necessary for the use of Congress—and for putting up a suitable apartment for containing them therein." The first books were ordered from England and shipped across the Atlantic in eleven hair trunks and a map case. The Library was housed in the new Capitol until August 1814, when—in a miniature replay of Julius Caesar's exploits in Alexandria—British troops invaded Washington and burned the building. The small congressional library of some three thousand volumes was lost in the fire. Another fire destroyed two-thirds of the collection in 1851. Unlike Alexandria, however, the Library of Congress has regrown, to approximately twenty-two million volumes today.

Returning to Ireland, the information explosion began to hit home in the Trinity College Library in the middle of the 19th Century. Fortunately, Omar's solution was not adopted. Instead, work started in 1835 on the production of a printed catalog, but by 1851 only the first volume, covering letters A and B, had been completed. The catalog was finally finished in 1887, but only by restricting the books that appeared in it to those published up to the end of 1872. Other libraries, however, were beginning to be faced with much bigger volumes of information. By the turn of the century, the Trinity College library had around a quarter of a million books, while the Library of Congress had nearly three times that number. Both were dwarfed by the British Museum, which at the time had nearly two million books, and the French National Library in Paris with over 2.5 million.

3 The Internet as a global information resource

Nearly a hundred years later, computer networks began and the floodgates really opened. As has often been observed, the real impact of computers is not so much in computation as it is in information and communication. The appearance of the personal computer two decades ago was a minor revolution. But what we have experienced in the past decade, with the advent of the Internet, the World Wide Web, the CD-ROM and DVD, is a major revolution in information and the communication of information. Large-scale information technology has finally worked its way into popular culture.

The Internet is the world's largest computer network—a network of networks, really—and one of its most long-standing and popular services is the

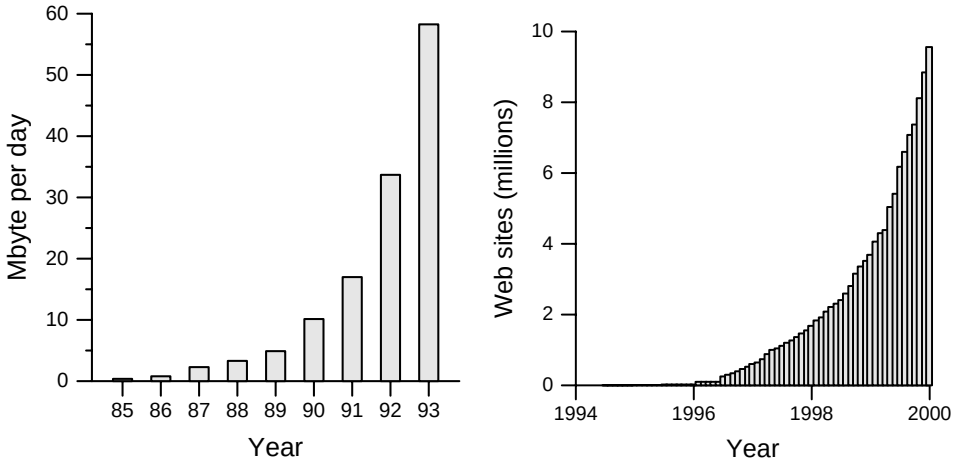


Figure 1: (a) Early growth of Internet news; (b) Growth of the Web

Usenet news service. This is a loose collection of news groups contributed by a huge user community, and it's free. To give an idea of the beginnings of the information explosion on computer networks, Figure 1a shows its initial growth in terms of the total number of megabytes contained in news articles per day: it doubled each year. In 1993 the volume of news broke 100 Mbs (roughly equivalent to 400 printed books) *per day*. In 1997, daily news broke 1 Gb, and the entire collection of news from before 1994—all the news represented by Figure 1a—was just three hours' worth at the then-current rate! The volume of news is no longer newsworthy: we cannot even locate annual figures to update the graph.

Even more alarming, though, is the rate of growth. At least until very recently—and perhaps even now—the number of articles, newsgroups, megabytes, users, and computers on the Internet have all been increasing exponentially since statistics began being collected in late 1984. Clearly, this cannot continue forever: there are some limiting factors. For example, early projections of the rate of growth of Internet users and the rate of growth of world population indicated that the former would overtake the latter in the year 2000![4] Inevitably things have slowed down somewhat: the number of electronic mail users in 1997 was 80 million and was projected to

[4]To put this ludicrous projection into perspective, it is said that half the world's population does not live within two hours' walk of a telephone.

grow (only!) tenfold by January 2001. Nevertheless, the number of Internet hosts has more than doubled every single year from 1982 (235 hosts) to the present day (72,400,000 hosts in 2000). When any exponential growth continues over a substantial period, mechanisms have to change, and the change generally occurs in a way that alters the nature of the questions we ask. For example, Internet Service Providers today allocate temporary host numbers to dial-in users: how should these be accounted for in the total host count?

Today's Internet wonder-child is the World Wide Web. It has become so pervasive so quickly that we tend to forget how recently it developed. Seven years ago, it was completely inconspicuous (150 hosts in mid-1993). Its growth is shown in Figure 1b. In 1998, there were estimated to be 320 million pages of information on the Web (Lawrence and Giles, 1998), and it continues to grow exponentially, doubling every six months. The average lifetime of a document is only 75 days, which accounts for the perceived unreliability of the medium as a serious information resource. To counter this, an organization called Internet Archive is aiming to archive the entire contents of the Web at regular intervals to preserve it for posterity (Kahle, 1997). Again, exponential growth forces mechanisms to change. A huge, and increasing, volume of information on the Web is now hidden behind portals such as virtual libraries or database systems, inaccessible to page-counting robots.

Much of the information on the Internet is ephemeral, trite, frivolous, and banal—and certainly does not compare with the information in the great libraries of the world. However, there are signs that something more serious is beginning to happen. Consider Project Gutenberg, whose goal is to encourage the creation and distribution of electronic text. Its aim is to have ten thousand electronic texts in distribution by the end of 2001. Conceived in 1971, the project's first achievement was an electronic version of the United States' *Declaration of Independence*, followed by the *Bill of Rights* and the *Constitution*. These were later followed by the Bible and Shakespeare—unfortunately, however, the latter was never released due to copyright restrictions. Like the Internet news, Project Gutenberg is a grass-roots phenomenon—in fact, the Internet is full of grass-roots phenomena. The amount of electronic text added to the Gutenberg collection doubles every year, with one book per month in 1991, two in 1992, four in 1993, and so on, so that the final goal should be reached in 2001. The project is still on schedule: currently (early 2000) there are 3,000 titles in the collection. Text is input by volunteers, each of whom can enter a book a year or even just one book in a lifetime. The project does not direct the volunteers' choice

of material; instead, people are encouraged to enter books they like, and to enter them in the manner in which they are comfortable. Quality control is likely to be a problem.

We noted earlier that in 1937, the renowned science fiction author H.G. Wells was promoting the concept of a "world brain" based on a permanent world encyclopedia (Wells, 1938):

> ... our contemporary encyclopedias are still in the coach-and-horses phase of development, rather than in the phase of the automobile and the aeroplane. Encyclopedic enterprise has not kept pace with material progress... the modern facilities of transport, radio, photographic reproduction and so forth are rendering practicable a much more fully succinct and accessible assembly of facts and ideas than was ever possible before.

It was supposed to give universal access to all human knowledge:

> Every university and research institution should be feeding it... its contents would be the standard source of material for the instructional side of school and college work, for the verification of facts and the testing of statements—everywhere in the world.

It is hard to resist pointing to the Internet as the beginning of a phenomenon that might broadly resemble a world encyclopedia. With its 72 million computers (in 2000), each equipped with, say, 1 Gb of storage, it has been described as the world's largest library. If just a quarter of one percent of this disk space were allocated for community use, the total space would amount to over 150 terabytes (150,000,000 Mb). It was estimated in 1975 that some 50,000,000 books had been published up to that time[5] (Gore, 1976). Even a tenth of 150 terabytes should be enough to accommodate a full-text database containing the text of all 50,000,000 books, compressed and indexed using standard techniques (Witten *et al.*, 1999)—and the remaining nine-tenths might provide space to add all the books published since 1975![6] The Alexandrian dream—and Wells's—would have been realized.

[5] On the assumption that half the people who ever lived are still alive, the total number of people who ever trod the planet up to 1975 is around 8,000,000,000, which means that on average one person in every hundred or two writes a book. That sounds reasonable—although it all depends what you mean by a "book."

[6] A well-known joke in the compression community observes that the Internet is so enormous that it can be used to compress any piece of text into just 13 bytes by the following strategem. Since any possible text must appear *somewhere* on the Internet, a 4-byte node address, plus a 5-byte disk address, plus a 4-byte character count, are enough to specify it fully.

4 What is a virtual library?

A virtual library is an organized collection of digital information. Like the classical libraries of Section 2, virtual libraries aim to provide uniform access to very large document collections. Like the Internet, they are global in scope and can transcend political and geographic boundaries. They combine the traditional library functions of collecting, classifying, and archiving information with the almost instant access and location independence that is the hallmark of modern computer networks. And since their raw material is digital in form, they support the kind of automated techniques for coping with the information explosion that we have seen in the previous section.

Ten definitions of the term "digital library" (here used synonomously with "virtual library") have been culled from the literature by Fox (1998), and their spirit is captured in the following brief characterization (Akscyn and Witten, 1998):

> *a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection.*

This definition gives equal weight to user (access and retrieval) and librarian (selection, organization and maintenance). Other definitions in the literature, emanating mostly from technologists, omit—or at best downplay—the librarian's role, which is unfortunate because it is the selection, organization, and maintenance that will distinguish virtual libraries from the anarchic mess that we call the World Wide Web. However, virtual libraries tend to blur what used to be a sharp distinction between user and librarian—because the ease of augmenting, editing, annotating and re-organizing electronic collections means that they will support the development of new knowledge *in situ*.

Virtual libraries are libraries without walls. But they do need boundaries. The very notion of a collection implies a boundary: the fact that some things are in the collection implies that others must lie outside it. And collections need a kind of presence, a conceptual integrity, that gives them cohesion and identity. Digital collections often present an appearance that is extremely opaque, a screen—typically a Web page—with no indication of what, or how much, lies beyond: whether a carefully-selected collection or a morass of worthless ephemera; whether half a dozen documents or many millions. At least physical libraries occupy physical space, present a physical appearance, and exhibit tangible physical organization. When standing on the threshold of a large bricks-and-mortar library one gains a sense of

presence and permanence that reflects the care taken in building and maintaining the collection inside. No-one could confuse it with a dung-heap! Yet in the virtual world the difference is not so palpable.

Thus we draw a clear distinction between a virtual library and the World Wide Web: the latter lacks the essential features of selection and organization that are central to the modern idea of a library. We would also like to make a distinction between a virtual library and a web *site*. Because extant virtual libraries invariably manifest themselves as web sites, the question arises whether any web site that provides a wealth of digital objects and provides appropriate methods of access and retrieval should be considered a "library." We believe not, not if the site achieves the facilities for access and retrieval by hand-crafted hypertext linkage structures. Inherent in the nature of libraries is maintainability: it should be easy to add new material to the library without having to rework linkage structures manually. To add new acquisitions to a physical library does not involve delving into the books and rewriting parts of them; similarly, it should be possible for new material to become a first-class member of a virtual library without any need for manual adjustment of the structures used for access and retrieval.

Creating a publicly-available virtual library presents interesting challenges, particularly if the library is intended for serious professional work rather than casual browsing. First, the raw material: the collection must comprise text that can be placed in the public domain. Second, the selection of material: although a huge amount of public-domain text is available on the Internet, its quality is extremely uneven and only a tiny fraction is appropriate for inclusion in a library. Third, the format of material: since most people prefer to do serious, sustained reading off-line rather than online, it is helpful if the collection resembles the traditional form of typeset pages rather than raw electronic text so that it can be printed for subsequent reading. Fourth, cataloging: appropriate bibliographic information in a usable format may be difficult to find and onerous to provide manually. Fifth, information retrieval: a uniform, easy-to-use, publicly-accessible interface is necessary.

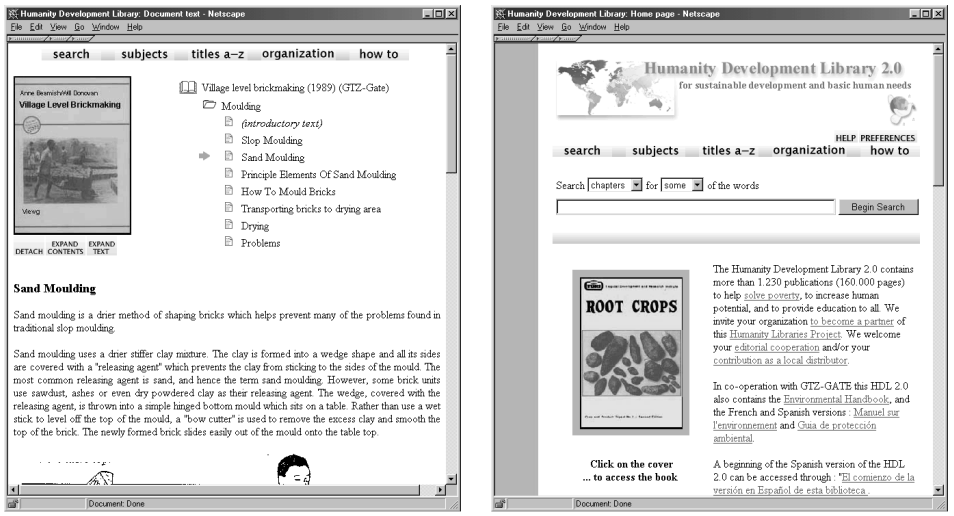
A wide spectrum of experimental virtual libraries have been created and are available on the World Wide Web. Some research projects focus on the underlying computer science technology, and are particularly concerned with standards and interoperability of libraries that are distributed in terms of both the location of the information itself and the organizational responsibility for collecting and cataloging that information. Other projects focus on collections, the digitizing and preservation of information objects. In some cases the library is intended to showcase particular national resources—for

example, historic, literary or artistic treasures—while in others it is intended to support access to information that is more international in character and oriented towards particular business or academic functions. Some projects provide full professional cataloging of the information in the library, others rely on catalog information donated by the institutions, or even the individuals, who supply the library material.

Cataloging tends to be a serious bottleneck in digital collections—as indeed it is in conventional library collections, as the Trinity College librarians discovered in the mid 19th Century. The professional judgements of trained and experienced librarians are in short supply, and cannot—in principle—cope with a world of exponentially growing information. This factor is somewhat mitigated by the sharing of catalog information that networked access makes possible: however, the responsibility for properly cataloged material rests with the individual library and is difficult to devolve.

More and more information today is becoming subsumed within the category of so-called "gray literature," ranging from the enormous volume of material that emanates from large international organizations like the United Nations down to the technical reports issued by individual university departments. The principal distinguishing feature of gray literature is that it lies outside the normal bookselling channels, which makes it more difficult to identify and acquire. Gray literature includes research reports, committee reports, conference papers, government reports, theses and dissertations, trade literature, and so on. It is particularly well suited to virtual library technology because, not being produced for profit, it is often released in electronic form which can be freely distributed on computer networks. However, the provision of formal cataloging information is a major obstacle to the creation of virtual libraries for gray literature. And without a title, author and subject database it seems hard to offer the searching facilities that are expected in physical libraries.

This is where automated tools for coping with the information explosion come in. If full text is available electronically, the search techniques explained in this book can be used to approximate the facilities offered by a conventional library catalog. And full-text searching of content can provide ways of locating information that are far in advance of those available in a conventional library. No back-of-the-book index can compare with a computerized text search—the brute power of the machine more than compensates for the human intelligence and effort that goes into a carefully-prepared index. Moreover, automatic agents can attempt to organize the information into clusters and categories. Methods of textual data mining can attempt to identify author and title information from the raw text, and



(a) (b)

Figure 2: (a) *Village Level Brickmaking*, (b) Humanity Development Library home page

perhaps identify references to other documents so that automatic hyperlinks can be made to them.

5 The Humanity Development Library

Figure 2a shows a book in the *Humanity Development Library*, a collection of humanitarian information put together by the Global Help Project to address the needs of workers in developing countries (`www.nzdl.org/hdl`). Because developing countries do not have ready access to the Internet, this collection is issued on a CD-ROM (as well as being available on the Web). The CD-ROM works on any version of the Windows operating system, and is easy to install—you need only insert it into the computer's CD-ROM drive and the system guides you through the installation instructions. The software provides a simple Web-browser interface on a standalone system. It also acts as a Web server, so if there are other computers on the same network (for example, in a school or hospital intranet), they automatically gain access to the collection too.

The Humanity Development Library CD-ROM, pictured in Figure 3a, is just one of several CD-ROMs that have been produced by humanitarian NGOs, including United Nations agencies such as UNESCO and the UN

(a) (b)

Figure 3: (a) Humanity Development Library CD-ROM; (b) Other humanitarian CD-ROMs

University; Figure 3b shows some others. The United Nations is greatly concerned about the imbalance in access to communication facilities. A statement on *Universal Access to Basic Communication and Information Services* issued by the UN's Administrative Committee on Coordination in 1997 remarks:

> We are profoundly concerned at the deepening mal-distribution of access, resources and opportunities in the information and communication field. The information technology gap and related inequities between industrialized and developing nations are widening: a new type of poverty—information poverty—looms. Most developing countries, especially the Least Developed Countries (LDCs) are not sharing in the communication revolution, since they lack:
>
> - affordable access to core information resources, cutting-edge technology and to sophisticated telecommunication systems and infrastructure;
> - the capacity to build, operate, manage, and service the technologies involved; policies that promote equitable public participation in the information society as both producers and consumers of information and knowledge; and
> - a work force trained to develop, maintain and provide the value-added products and services required by the information economy.

> We therefore commit the organizations of the United Nations system to assist developing countries in redressing the present alarming trends.

The existence of this information on the Web and as CD-ROMs illustrates the phenomenal improvements that virtual libraries make to the accessibility of information—and, as we saw earlier, accessibility has been a touchstone of library service since the very beginning of libraries. The Humanity Development Library contains 1250 books and magazines that weigh 340 Kg and cost US$20,000 to reproduce, yet it fits easily on a single CD-ROM that weighs almost nothing and costs US$6.

The book in Figure 2a might have been reached by a directed full-text search, or by browsing one of a number of access structures, or by clicking on one of a gallery of images. On opening the book, which is entitled *Village Level Brickmaking*, a picture of its cover appears at the top, beside a hierarchical table of contents. In the figure, the reader has drilled down into a chapter on *moulding* and a subsection on *sand moulding*, whose text appears below. Readers can expand the table of contents from the section to the whole book; and expand the text likewise (which is very useful for printing). The ever-present picture of the book's cover gives a feeling of physical presence and a constant reminder of the context.

Readers can browse the collection in several different ways, as determined by the editor who created it. Figure 2b shows the collection's home page, at the top of which (underneath the logo) is a bar of five buttons that open up different access mechanisms. A subject hierarchy provides a tree-structured classification scheme for the books. Book titles appear in an alphabetical index. A separate list gives participating organizations and the material that they contributed. A "how-to" list of helpful hints, created by the collection's editor, allows a particular book to be accessed from brief phrases that describe the problems the book addresses. However a book is reached, it appears in the standard form illustrated in Figure 2a, along with the cover picture to give a sense of presence. The different access mechanisms help solve the librarian's dilemma of where to shelve books (Mann, 1993): each one appears on many different virtual shelves, shelves that are organized in different ways.

Full-text search of titles and entire documents provide important additional access mechanisms. The search engine that we use, MG (Witten *et al.*, 1999), supports searching over the full text of the document—not merely a document surrogate as in conventional computer-based library retrieval systems. User feedback from an earlier version of this collection indicated

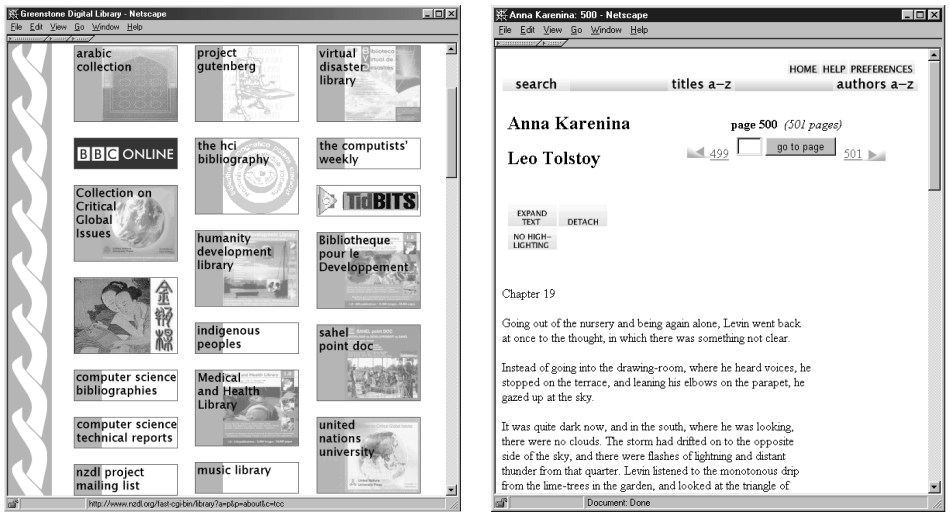
that Boolean searching was more confusing than helpful for the targeted users. Previous research suggests that difficulties with Boolean syntax and semantics are widespread, and transaction log analysis of several library retrieval systems indicates that by far the most popular Boolean operator is AND; the others are rarely used. For all these reasons, the interface default for this collection is ranked queries. However, to enable users to construct high-precision conjunctive searches where necessary, selecting "search ... for *all* the words" in the query dialog produces the syntax-free equivalent of a conjunctive query.

Just as libraries display new acquisitions or special collections in the foyer to pique the reader's interest, this collection's home page (Figure 2b) highlights a particular book that changes every few seconds: it can be opened by clicking on the image. This simple display is extraordinarily compelling. And just as libraries may display a special book in a glass case, open at a different page each day, a "gallery" screen can show an ever-changing mosaic of images from pages of the books, remarkably informative images that, when clicked, open the book to that page. Or a scrolling "Times Square" display of randomly selected phrases that, when clicked, take you to the appropriate book. The possibilities are endless.

The Humanity Development Library is a focused collection of 1250 books—miniscule by library standards, but nevertheless comprehensive within the targeted domain. It contains 53,000 chapters, 62 million words, 32,000 pictures. Although the text occupies 390 MB, it compresses to 102 MB and the two indexes—for titles and chapters respectively—compress to less than 80 MB. The images (mostly in PNG format) occupy 290 MB. Associated files bring the total size of the collection to 505 MB. It fits comfortably on a CD-ROM, and even if there were twice as much text, and the same images, it would still fit, along with all the necessary software. A single DVD-ROM would hold a collection twenty times the size—still small by library standards, but immense for a fully portable collection.

6 Other virtual library collections

The Humanity Development Library is just one of about two dozen publicly-available collections produced by the New Zealand Digital Library (NZDL) project and listed on the project's home page (`www.nzdl.org`), part of which is shown in Figure 4a. This project aims to develop the underlying infrastructure for virtual libraries and provide example collections that demonstrate how it can be used. Figure 4a illustrates the wide range of collections.



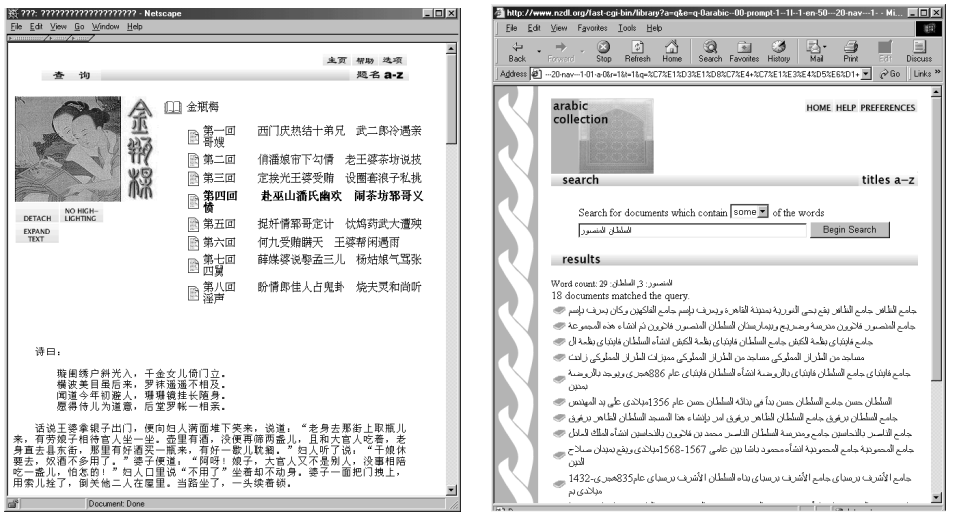
(a) (b)

Figure 4: (a) Some virtual library collections, (b) Reading Tolstoy's *Anna Karenina*

The library is international: there are interfaces in English, Maori, French, German, Arabic, and Chinese, and collections have been produced in all these languages. Virtual libraries are particularly empowering for the disabled, and there is a text-only version of the interface intended for visually impaired users.

The editors of the Humanity Development Library have gone to great lengths to provide a rich set of access structures. However, this is a demanding, labor-intensive task, and most collections are not so well organized. The basic access tool in the NZDL is full-text searching, which is available for all collections and is provided completely automatically when a collection is built. Some collections allow, in addition, traditional catalog searching based on author, title, and keywords, and full-text search within abstracts. Our experience is that while the user interface is considerably enhanced when traditional library cataloging information is available, it is often prohibitively expensive to create formal cataloging information for electronically-gathered collections. With appropriate indexes, full-text retrieval can be used to approximate the services provided by a formal catalog.

The historically first collection in the New Zealand Digital Library is the *Computer Science Technical Reports* collection, now containing 46,000 reports—1.3 million pages, half a billion words—extracted automatically



(a) (b)

Figure 5: (a) Reading a Chinese book; (b) Searching an Arabic collection

from 34 GB of raw PostScript. There is no bibliographic or "metadata" information: we have only the contents of the reports (and the names of the FTP sites from which they were gathered). Many are Ph.D. theses which would otherwise be effectively lost except to a miniscule community of cognoscenti: full-text search reaches right inside the documents and makes them accessible to anyone looking for information on that topic.

As well as the simplified searching interface for the Humanity Development Library described above, users can choose a more comprehensive query interface (via a *Preferences* page). Case-folding and stemming can be independently enabled or disabled, and full Boolean query syntax is supported as well as ranked queries. Moreover, in the *Computer Science Technical Reports* searches can be restricted to the first page of reports, which approximates an author/title search in the absence of specific bibliographic details of the documents. Although this is a practical solution, the collection nevertheless presents a raw, unpolished appearance compared with the Humanity Development Library, reflecting the difference between a carefully-edited set of documents, including hand-prepared classification indexes and other metadata, and a collection of information pulled mechanically off the Web and organized without any human intervention at all.

There are several collections of books, including, for example, the English books entered by the Gutenberg project. Figure 4b shows a user reading a

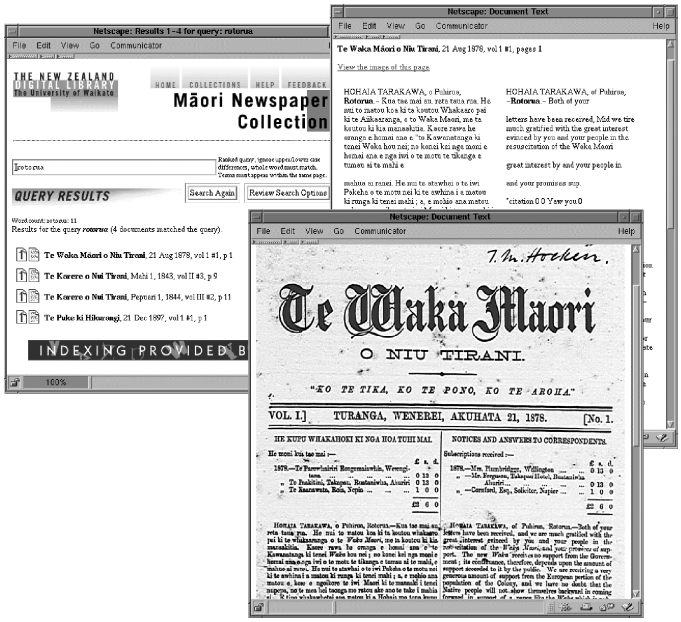


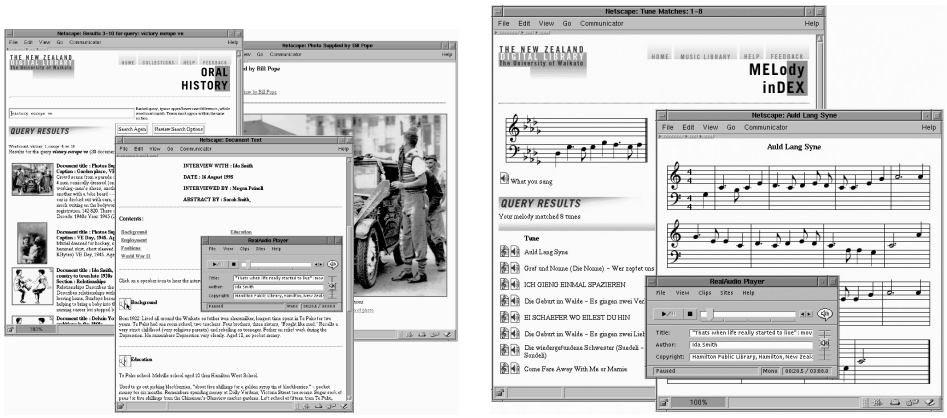
Figure 6: Searching the *Historic New Zealand newspapers* collection

book in this collection: Tolstoy's *Anna Karenina*. The Gutenberg collection does not specify a specific hierarchical structure for its documents: thus, a page number, next- and previous-page function, and page selection box are included rather than the hierarchical structure illustrated in Figure 2a.

Figure 5a shows a book in a collection of classical Chinese literature. The full text was available on the Web; we automatically extracted the section headings to provide the table of contents visible at the upper right, and scanned the book's cover to generate the cover image. One can perform full-text search on the complete contents or on section headings alone, using the Chinese language (of course your Web browser must be set up correctly to work in Chinese). There is also a browsable list of book titles. Figure 5b illustrates searching an Arabic collection: note that the search terms are entered in the Arabic language using the Web browser's foreign-language facility.[7]

An expressly bilingual collection of *Historic New Zealand Newspapers* contains issues of forty newspapers published between 1842 and 1933 for a Maori audience. Collected on microfiche, these constitute 12,000 page im-

[7] Normally one would access this with an Arabic interface; we have selected the English interface for the reader's convenience.



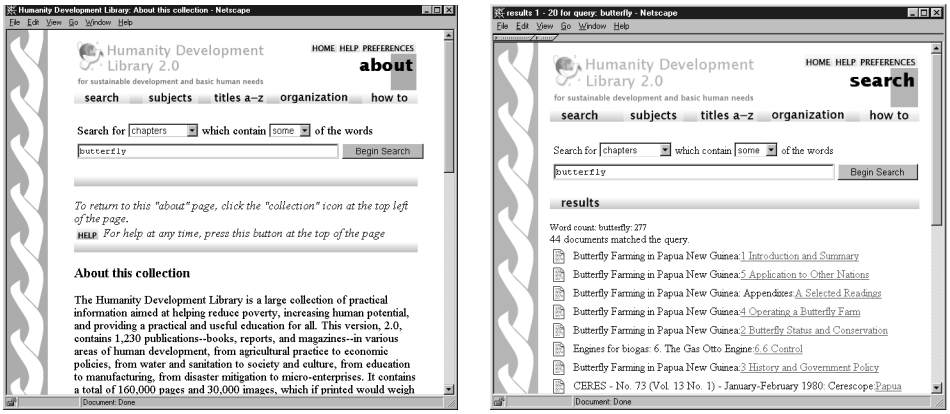
(a) (b)

Figure 7: (a) Results of a query about VE Day from the Oral History Collection; (b) Using the Music Library to find *Auld Lang Syne*

ages. Although they represent a significant resource for historians, linguists and social scientists, their riches remain largely untapped because of the difficulty of accessing, searching and browsing material in unindexed microfiche form. Figure 6 shows the parallel English-Maori text retrieved from the newspaper *Te Waka Maori* of August 1878 in response to the query *Rotorua*, a small town in New Zealand. Searching is carried out on electronic text produced using OCR; once the target is identified, the corresponding page image can be displayed.

Some collections contain more than just text and images. For example, the *Oral History Collection* includes recorded speech. Summary transcripts of taped interviews, photographs that illustrate episodes mentioned in the tapes and appropriate timing information are utilized to retrieve audio extracts and pictures that match a given query. Figure 7a shows some results from a query about VE Day. The small gray panel controls audio playback of a sound bite or relevant section from a participant's interview.

A technically more significant multimedia capability is demonstrated by a retrieval engine for music. Based on a novel scheme for searching musical melodies, this matches sung (or hummed) input to a database of tunes. Figure 7b shows the response when a user sang the first eight notes of *Auld Lang Syne* as a query. The transcribed input appears at the top left; titles of similar items, ranked according to how closely they match the query, appear below. Any of the tunes may be selected for audio replay or visual display; one appears in the front window. A database of nearly



(a) (b)

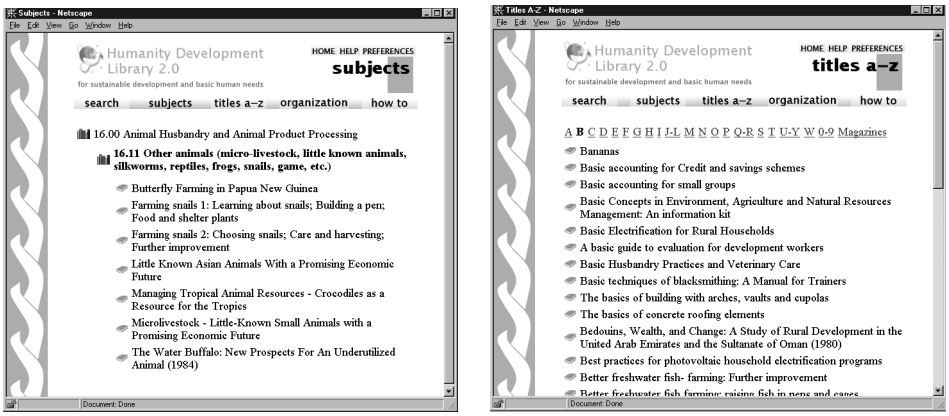
Figure 8: Searching the Humanity Development Library: (a) search page, (b) search results

ten thousand international folk tunes (half a million notes) from various countries (North America, Ireland, Britain, Germany, and China) is based on this retrieval system. Tunes from each country can be searched separately or in combination. This collection demonstrates the need to encompass radically different search regimes within a single unifying architecture.

7 The Greenstone software

All these collections are created using the Greenstone software developed by the NZDL project (Witten *et al.*, 2000). Information collections built by Greenstone combine extensive full-text search facilities with browsing indexes based on different metadata types. There are several ways for users to find information, although they differ between collections depending on the metadata available and the collection design. You can *search for particular words* that appear in the text, or within a section of a document, or within a title or section heading. You can *browse documents by title*: just click on the displayed book icon to read it. You can *browse documents by subject*. Subjects are represented by bookshelves: just click on a shelf to see the books. Where appropriate, documents come complete with a table of contents (constructed automatically): you can click on a chapter or subsection to open it, expand the full table of contents, or expand the full document.

An example of searching is shown in Figure 8, where the Humanity Development Library is being searched for chapters that contain the word



(a) (b)

Figure 9: Browsing the Humanity Development Library (a) by subject, (b) by title

butterfly. The first part of the Figure shows the user entering the search, and the second shows the results obtained. Fourty-four chapters contain the search term, of which the first ten are shown in Figure 8b. Six of these are chapters of the book *Butterfly Farming in Papua New Guinea*. Clicking on one of these will show the book in the format illustrated in Figure 2a, opened at the chapter in question.

In Figure 9a, the same collection is being browsed by subject: by clicking on the bookshelf icons the user has discovered an item under *Section 16, Animal Husbandry*. Pursuing an interest in butterfly farming, the user selects a book by clicking on its icon. The book will then be shown in the same format as before. Figure 9b shows the collection being browsed by title for books whose title begins with the letter *B*.

As this example illustrates, an important distinction is made between *searching* and *browsing*. Searching is full-text, and—depending on the collection's design—the user can choose between indexes built from different parts of the documents, or from different metadata. Some collections have an index of full documents, an index of sections, an index of paragraphs, an index of titles, and an index of section headings, each of which can be searched for particular words or phrases. Browsing involves data structures created from metadata that the user can examine: lists of authors, lists of titles, lists of dates, hierarchical classification structures, and so on. Data structures for both browsing and searching are built according to instructions in a configuration file, which controls both building and serving the

collection.

Rich browsing facilities can be provided by manually linking parts of documents together and building explicit indexes and tables of contents. However, manually-created linkages become difficult to maintain, and often fall into disrepair when a collection expands. The Greenstone software takes a different tack: it facilitates *maintainability* by creating all searching and browsing structures automatically from the documents themselves. No links are inserted by hand. This means that when new documents in the same format become available, they can be added automatically. Indeed, for some collections this is done by processes that wake up regularly, scout for new material, and rebuild the indexes—all without manual intervention.

Collections comprise many documents: thousands, tens of thousands, or even millions. Each document may be hierarchically organized into *sections* (subsections, sub-subsections, and so on). Each section comprises one or more *paragraphs*. Metadata such as author, title, date, keywords, and so on, may be associated with documents, or with individual sections of documents. This is the raw material for indexes. It must either be provided explicitly for each document and section (for example, in an accompanying spreadsheet) or be derivable automatically from the source documents. Metadata is converted to Dublin Core and stored with the document for internal use.

All collections created with Greenstone contain a brief statement of purpose and coverage, which articulates the principles governing what material is included in the collection, and a brief explanation of how the collection is organized. The beginning of these can be seen, for example, on the "about" page of the Humanity Development Library in Figure 8a.

In order to accommodate different kinds of source documents, the software is organized so that "plugins" can be written for new document types. Plugins exist for plain text documents, HTML documents, email documents, and bibliographic formats. Word documents are handled by saving them as HTML; PostScript ones by applying a preprocessor (Nevill-Manning *et al.*, 1998). Specially written plugins also exist for proprietary formats such as that used by the BBC archives department. A collection may have source documents in different forms: it is just a matter of specifying all the necessary plugins. In order to build browsing indexes from metadata, an analogous scheme of "classifiers" is used: classifiers create indexes of various kinds based on metadata. Source documents are brought into the Greenstone system through a process called *importing*, which uses the plugins and classifiers specified in the collection configuration file.

The international Unicode character set is used throughout, so documents—

and interfaces—can be written in any language. Colections have so far been produced in a wide variety of languages; the NZDL web site provides numerous examples. Collections can contain text, pictures, and even audio and video clips. Compression technology is used to ensure best use of storage (Witten *et al.*, 1999). Most non-textual material is either linked to textual documents or accompanied by textual descriptions (such as photo captions) to allow full-text searching and browsing. However, the architecture permits the implementation of plugins and classifiers even for non-textual data.

The system includes an "administrative" function whereby specified users can examine the composition of all collections, protect documents so that they can only be accessed by registered users on presentation of a password, and so on. Logs of user activity are kept that record all queries made to every Greenstone collection (though this facility can be disabled).

Although primarily designed for Internet access over the World Wide Web, collections can be made available, in precisely the same form, on CD-ROM—as with the Humanity Development Library in Section 5. In either case they are accessed through any Web browser. Greenstone CD-ROMs operate on a standalone PC under Windows 3.X, 95, 98, and NT, and the interaction is identical to accessing the collection on the Web—except that response is faster and more predictable. The requirement to operate on early Windows systems is a significant practical impediment to the software design, but is crucial for many users—particularly those in underdeveloped countries seeking access to humanitarian aid collections. If the PC is connected to a network (intranet or Internet), a custom-built Web server provided on each CD makes exactly the same information available to others through their standard Web browser. The use of compression ensures that the greatest possible volume of information can be packed on to a CD-ROM.

The collection-serving software operates under Unix and Windows NT, and works with standard Web servers. A flexible process structure allows different collections to be served by different computers, yet be presented to the user in the same way, on the same Web page, as part of the same virtual library (McNab *et al.*, 1998). Existing collections can be updated and new ones brought on-line at any time, without bringing the system down; the process responsible for the user interface will notice (through periodic polling) when new collections appear and add them to the list presented to the user.

8 Conclusion

Today, we stand on the threshold of a new era of virtual libraries. Conceived by visionary thinkers and fertilized with resources by today's politicians, they are undergoing a protracted labor and birth. Virtual libraries are arriving just at the time when our physical libraries are cracking under the stress of continued exponential growth. They will be different from the World Wide Web: libraries are focused collections, and it is the act of selection that gives them focus. For many practical reasons (including copyright, and the physical difficulty of digitization), virtual libraries will not vie with archival national collections, not in the foreseeable future. Their role is in specialist, targeted collections of information.

Established libraries of printed material have sophisticated and well-developed human and computer-based interfaces to support their use. But they are not well integrated for working with computer tools: a bridging process is required. Information workers can immerse themselves physically in the library, but they cannot take with them their tasks, tools, and desktop workspaces. The virtual library will be different: we will work "inside" it in a sense that it totally new.

But even for a focused collection, creating a high-quality virtual library is a highly labor-intensive process. To provide the richness of access and inter-connection that makes a virtual library comfortable requires enormous editorial effort. And when the collection changes, maintenance becomes an overriding issue. Fortunately, techniques of text mining are emerging that offer the possibility of automatic identification of semantic items from plain text. Carefully-constructed user interfaces can take advantage of the information that they generate to provide a library experience that is qualitatively different from a physical library—not just in access and convenience, but in terms of the quality of browsing and information accessibility. Tomorrow, virtual libraries will put the right information right at your fingertips.

We began this essay by recounting the Alexandrian principle that the ideal library should *universal*: it should contain all the world's recorded knowledge. We then discussed the principle of *accessibility*: libraries are for all. In Hegelian terms, these two principles together constitute the *thesis*, the proposition to be maintained. Combine them with a third ingredient, the sustained exponential growth of recorded knowledge, and the contradiction leaps out—how can one collect so much stuff and still make it accessible? This is the *antithesis*: physical books occupy space, their organization takes clerical effort, making them accessible requires a manually-operated service; continued exponential growth is insupportable.

Our argument is that virtual libraries are emerging as the *synthesis*, a higher conception that reconciles thesis and antithesis by transcending both. It is the physical nature of the material that causes the contradiction. Defining a virtual library as a collection of digital objects, along with methods for access and retrieval, and for selection, organization, and maintenance, provides a way out. Of course, for full scalability, access, retrieval, organization, and maintenance must be accomplished automatically, without manual intervention. And we have shown glimpses of how this can be done. The Humanity Development Library is a virtual collection that makes physical books widely accessible. The New Zealand Digital Library gives many examples of different kinds of virtual collections. The Greenstone software makes it easy to build such collections from existing literature.

Today we stand on the threshold of a virtual library that will bear comparison with, and ultimately supersede, the great libraries of history. To realize their full potential, universal virtual libraries will demand an unprecedented degree of international cooperation. They will strain to breaking point current social mechanisms such as copyright and intellectual property protection (ACM, 1999). Coping with continual growth will create enormous problems of review and academic authority. But if these challenges are met, future digital libraries will dramatically improve access to the world's knowledge—not just snapping the links of the chained book, but vaporizing the chain and dematerializing the book. They will also act as "collaboratories" out of which new knowledge is crafted and refined by widely-distributed teams and organizations—knowledge that right from conception is fully interconnected with previous work. The librarians of Alexandria would have been just as excited by this prospect as we are.

Acknowledgements

Many thanks to members of the New Zealand Digital Library project, whose work is described in this paper, particularly David Bainbridge, George Buchanan, Stefan Boddie, Rodger McNab and Bill Rogers who built the Greenstone system; Craig Nevill-Manning who worked on an early version; and Mark Apperley, Carl Gutwin, Steve Jones, Te Taka Keegan, Malika Mahoui and Gordon Paynter for their help and support. Sally Jo Cunningham made greatly appreciated comments and constructive suggestions on a draft of this paper.

References

ACM (1999) *Intellectual property in the age of universal access.* ACM Press.

Akscyn, R.M. and Witten, I.H. (1998) "Report on First Summit on International Cooperation on Digital Libraries." `ks.com/idla-wp-oct98`.

Bowker, R.R. (1883) "The work of the nineteenth-century librarian for the librarian of the twentieth." *Library Journal 8*: 247–250; September-October; quoted by Crawford and Gorman (1995).

Bush, V. (1947) "As we may think." *The Atlantic Monthly*, Vol. 176, No. 1, pp. 101–108.

Clarke, I. (1999) "A distributed decentralised information storage and retrieval system." M.Sc. Thesis, Division of Informatics, University of Edinburgh.

Crawford, W. and Gorman, M. (1995) *Future libraries: dreams, madness, and reality.* Americal Library Association, Chicago.

Follett, B. (1993) "Joint Funding Council's Library Review Group: Report," *Higher Education Funding Council for England*, Bristol, U.K.

Fox, E. (1998) "Digital library definitions." `ei.cs.vt.edu/ fox/dlib/def.html`.

Gore, D. (Editor) (1976) *Farewell to Alexandria: Solutions to space, growth, and performance problems of libraries.* Greenwood Press, Westport, Connecticut.

Kahle, B. (1997) "Preserving the internet." *Scientific American 276*(3): 82–83; March.

Kruse, R. (1994) "Human skin book." Web Posting to *Rare Books and Special Collections Forum*, 14 February 1994.

Lawrence, S. and Giles, C.L. (1998) "Searching the World Wide Web." *Science 280*: 98–100; April.

Lesk, M. (1997) *Practical digital libraries: Books, bytes, and bucks.* Morgan Kaufmann, San Francisco.

Licklider, J.C.R. (1960) "Man-computer symbiosis." *IRE Trans HFE-1*, pp. 4–11.

Mann, T. (1993) *Library research models.* Oxford University Press, NY.

McNab, R.J., Witten, I.H. and Boddie, S.J. (1998) "A distributed digital library architecture incorporating different index styles." *Proc IEEE Advances in Digital Libraries*, Santa Barbara, CA, pp. 36–45.

Nevill-Manning, C.G., Reed, T., and Witten, I.H. (1998) "Extracting text from PostScript." *Software—Practice and Experience*, Vol. 28, No. 5, pp. 481–491.

Thompson, J. (1977) *A history of the principles of librarianship.* Clive Bingley, London.

Wells, H.G. (1938) *World brain.* Doubleday, New York.

Witten, I.H., Moffat, A., and Bell, T.C. (1999) *Managing Gigabytes: Compressing and indexing documents and images.* Morgan Kaufmann, San Francisco.

Witten, I.H., McNab, R.J., Boddie, S.J. and Bainbridge, D. (2000) "Greenstone:

a comprehensive open-source digital library software system." *Proc ACM Digital Libraries*, San Antonio, TX.